

Release Notes

Dachuan Zhang
Matthew Quaglia
Serena (Jiahui) Zeng
Sri Lakshmi Tirupathamma Manduri

Overview

Our project, **The Snow and Ice Removal Prioritizer** aims to provide a platform to help city officials determine areas of a city to prioritize snow and ice removal. Additionally, it can be used to determine areas of the city that could benefit from lower-risk infrastructure or a higher presence of emergency response personnel.

Data sources

We utilized multiple data sources. Our main data sources are the US Accidents (2016-2023) dataset from Kaggle (Moosavi, *US accidents (2016 - 2023)* 2023). It contains data pertaining to car accidents on a national level. Our project focuses mainly on the Massachusetts traffic accident data. This project also utilizes some portions of the data from the IMPACT project data available from the Massachusetts government (*MassDOT: Crash Data Portal*).

Architecture

The project uses data selected from our primary and secondary datasets. It takes inputs from the user to generate other needed features of the model to provide an estimate of traffic accidents likelihood of a certain area. Locations are grouped based on the first three decimal points of latitude and longitude.

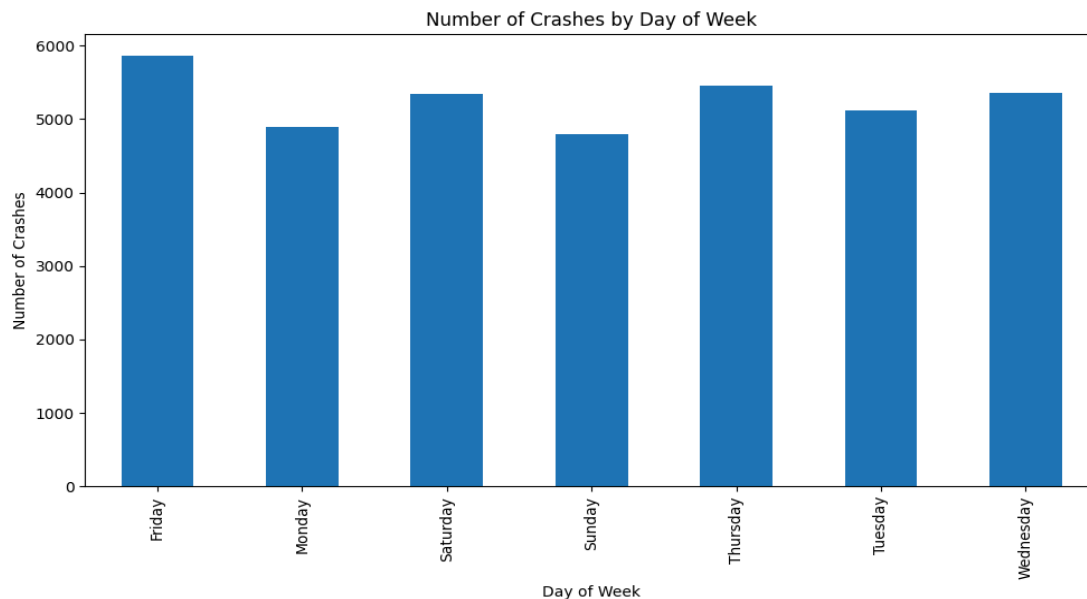
Deliverables and Outputs

The primary output of the interface, found at the bottom of the page, is a prioritized list of areas in a city based on the average likelihood of accidents throughout a day based on the city, road surface condition, and date. Additional outputs include interactive tables and maps based on historical accident data.

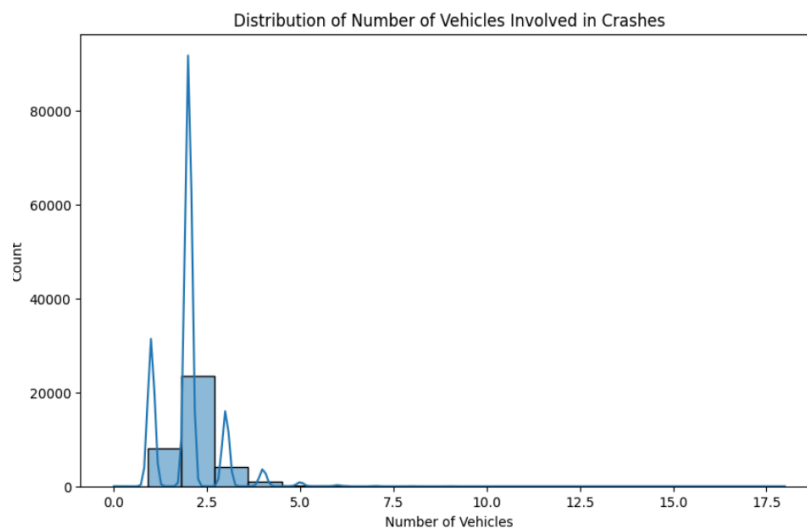
The output aims to help government officials determine what areas are most crucial in terms of reducing dangerous road surface conditions. It also gives information regarding accident rates and other information throughout Massachusetts cities to help government officials determine what areas could benefit from solutions such as safer infrastructure or traffic calming tools.

Exploratory Data Analysis (EDA)

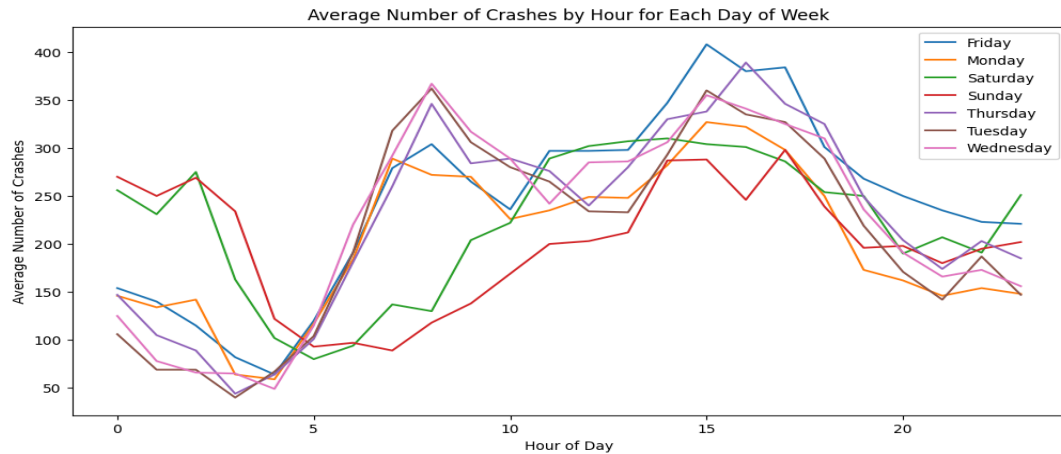
The figure below exhibits crashes with respect to the day of the week. You can see that Friday has the highest occurrence rate of accidents.



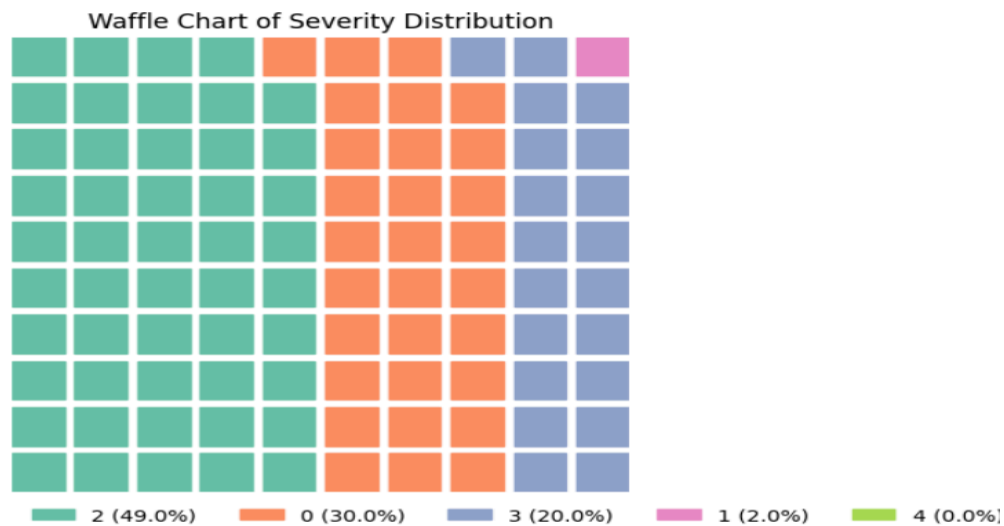
The figure below shows the number of vehicles involved in an accident. You can see that generally two cars are involved, and accidents with more than four vehicles are especially uncommon.



The figure below exhibits at what time of the day accidents often occur. On Fridays it is apparent that many accidents occur around 2 pm EST.



The figure below shows us the severity of an accident when it occurred. Most often, it is a severity of 2.



The figure below shows us that accidents are occurring more often during Dry conditions compared to wet conditions.



In the end, we designed an interface based on Dash for users to predict the coordinates to be prioritized for snow and ice removal. This application also offers other data exploration features such as an interactive data table for users to overview and filter data, a scatter mapbox for

visualizing the crash locations, and a choropleth map to visualize the count of accidents occurring at a county level.


Below is the table featured in our interface. It can be filtered based on any of the columns, allowing users to dig deeper in the underlying data of our model.

\$Start_Time	\$Severity	\$Street	\$City	\$County	\$State	\$Weather_Condition
filter data...						
2016-11-30 15:37:19	2	Fairview Ave	Chicopee	Hampden	MA	Light Rain
2016-11-30 16:14:24	2	Overbrook Dr	Wellesley	Norfolk	MA	Rain
2016-11-30 16:02:41	3	I-95 N	Lexington	Middlesex	MA	Rain
2016-11-30 14:12:49	4	MA-128 S	Burlington	Middlesex	MA	Light Rain
2016-11-30 16:00:47	3	Yankee Division Hwy S	Peabody	Essex	MA	Light Rain
2016-11-30 17:03:52	3	I-93 S	Medford	Middlesex	MA	Rain
2016-11-30 17:18:00	2	Washington St	Quincy	Norfolk	MA	
2016-11-30 17:33:47	2	S Main St	Andover	Essex	MA	Light Rain
2016-11-30 17:57:43	3	I-93 N	Dorchester	Suffolk	MA	Rain
2016-11-30 18:34:30	2	S Main St	Haverhill	Essex	MA	Light Rain
2016-11-30 18:43:32	2	Main St	Tewksbury	Middlesex	MA	Light Rain
2016-11-30 18:46:15	3	Fellsway W	Medford	Middlesex	MA	Rain
2016-11-30 19:05:51	2	Forest St	Medfield	Norfolk	MA	Overcast

<< < 1 / 4769 > >>

Instructions and Caveats for Use

After downloading County.json and all csv files, the application can be run with the script, app.py. Due to complexities in running the interface due to downloads and dependencies, we have provided a link to a Demo of the interface for those unable to run locally on their machine. Our next steps in this project are to run the interface on a cloud-based site for a simpler way to access the tool. See the link below to view the Demo:

 Demo.mp4

Frequently Asked Questions

What is the primary purpose of the interface?

The primary purpose of the interface is to determine what areas of a city should be prioritized for snow and ice removal. Due to the nature of the model, the interface results can give insights into similar use-cases, such as determining what areas should have a higher presence of emergency-response personnel.

What inputs does the user need to specify to get the final output?

The inputs of the model are City, Date, and Road Surface Condition.

What is the output of the interface? How can the user interpret the output?

The output of the interface is a ranked list of areas (assigned by latitude and longitude points to the third decimal point), ordered by predicted severity, which can be considered an average throughout the day of accident severity.

What does severity mean?

Severity is a score from 0 to 4. Severity indicates if an accident is likely to occur, and if it does occur what the damage will look like in terms of vehicle damage and injury. A severity of 0 indicates that an accident is unlikely to occur. A severity of 4 indicates that an accident is likely to occur with high damage to vehicles with severe and potentially fatal injuries likely.

What level does the interface consider? Can this be modified depending on the use case?

The interface considers a City-level model. Currently, this cannot be changed. Future iterations of this project could expand to a County-level or be reduced to a Neighborhood-level.

How to get more information

To get a deeper understanding of the data involved in the project, follow the links below:

Primary Datasource

[US Accidents \(2016-2023\)](#)

Secondary Datasource

Data Query and Visualization can be used to pull specific information about Massachusetts car accident data

[IMPACT \("A tool for researching crash-related data in Massachusetts"\)](#)

Methodological Appendix

Building and Training The Algorithm

To train our model, we first clean and combine the two datasets we are pulling data from pertaining to car accidents that have occurred in Massachusetts along with an auto-generated dataset pulling in data for accidents that did not occur. The primary dataset is US Accidents (2016-2023) found on Kaggle, which is chunked to only extract Massachusetts data. The secondary dataset, pulled from the IMPACT portal from MA gov, is used to pull in road surface conditions based on coordinates and datetime values. Columns with null values are dropped or imputed based on the amount of null values found and the potential of accurately filling in these values, such as Wind Chill, which can be generally assumed to be the same as temperature when the column is found to be null. To generate non-accident data, a function is created called `generate_random_Boston_coordinates`, which randomly selects the specified number of coordinates that fall within Boston. Infrastructure and weather-related columns are set with placeholder values initially, and they are later imputed with values from the primary dataset.

After the cleaning, implementing, and merging of the three datasets (primary, secondary, and auto-generated), the variables are encoded to be trained on the models. The data is initially run with all relevant columns using logistic regression, multiple-linear regression, and random forest regression. The models are designed to predict Severity. Severity is considered on a scale of 0-4, with 0 implying that no accident will occur and 4 implying that a very serious accident may occur that will result in severe and potentially fatal injuries. In general, an accident with notable damage but without serious injuries will result in a severity of 2. Upon initial training, random forest regression is the selected model. We use feature importance along with our own judgment to determine what columns are most relevant to pull in. The final model considers road surface condition, latitude, longitude, month, day, hour, minute, second, and several infrastructure columns including crossing, amenity, station, traffic signal, railway, give way, junction, and stop. The model runs with 100 estimators. The mean square error of the model is 0.101294 and the R^2 score is 0.919192.

Combining interface and model

The data frame that is cleaned and combined (both accident and non-accident data) is worked on for different latitudes and longitudes to generate city names in Massachusetts. For the prototype currently available, the only option is Boston, but it is designed to be able to easily add other cities in the future with some additional work.

User input:

- City
- Date
- Road Surface Conditions

Based on this user input, we have to generate a dataframe and populate True or False values for columns: 'Amenity', 'Crossing', 'Station', 'Traffic_Signal', 'Railway', 'Give_Way', 'Junction', 'Stop' by training the dataframe using the random forest model and testing it with the latitude and longitude of the data frame created based on city given in user input.

After the data frame is created with True or False values for different columns, we use the trained random forest model in the previous section of '*Building and Training The Algorithm*', which is tested on the created dataframe for the severity column.

API for future improvements and extra uses

The TomTom API (*Maps and location technology*) aims to provide extra information to give the traffic flow speed of certain areas. It can be incorporated into the interface to provide extra information to help users make further insights and more specifically prioritize our final output, the ranked list, ranking areas with more traffic higher than areas with equal severity but lower traffic rates.

Comparison with a similar project

This portion of the documentation will compare the interface we created with a similar project implemented for the city of Boston, Vision Zero. As stated at the top of the site page, “Vision Zero Boston is our commitment to focus the City’s resources on proven strategies to eliminate fatal and severe traffic crashes in the City by 2030” (*Vision Zero* 2018). Similarly to our project, the Vision Zero site allows for some interesting and interactive visualizations. Specifically, it displays Safety Concerns Map, Crash Map, and Vision Zero Dashboard. The Crash Map is especially interesting, as we have a similar map on our page showcasing historical data. An interesting feature that would be insightful to add in future iterations of our project is the ability to filter based on Pedestrian, Bicyclist, or Motor vehicle crash. The Vision Zero Dashboard displays interesting bar charts allowing the user to select one or multiple years to compare historical data. While we have several maps and a table where similar information can be found, it would be helpful to add automated visuals such as this to our interface so that the user can get further insights into the data. On the Safety Concerns Map, the user can pinpoint an area of the city where they believe an improvement can be made that may make the area safer for drivers, pedestrians, and cyclists. It would be useful to see something like this in our interface specifically for concerns regarding areas that users notice tend to be especially treacherous during inclement weather, specific infrastructure that feels unsafe during periods of snow, and things of this nature. Overall, Vision Zero is a very insightful resource for those interested in car accident data that inspires a great deal of improvements and functionalities that could be implemented in future iterations of our tool

What We Learned During The Project

Initially, our project idea was very broad. We wanted to create an interface that would predict the accident likelihood at a specific location. While the idea had potential, it was a bit too general. It made more sense to focus on a specific-use case. This feedback we received about our initial idea helped us understand the importance of focusing on a specific use-case. Generally, we learned that it is important to make a tool that has a clear audience. While our original idea would have interesting outputs, it was difficult for us to explain who would benefit from this interface and what they would be using it for. Honing in on a specific example allowed us to truly put ourselves in the shoes of our audience and think through how they can use the interface on a day-to-day basis. We were able to make visuals that tell the user historical information specific to road surface conditions and accidents over time. Most importantly, we were able to display concise results from our model that give the user direct answers to the specific question: Where should this city focus on snow and ice removal on this date?

Additionally, we learned the importance of building multiple pieces of the project simultaneously rather than doing so in the order that the project runs. During the previous project that the majority of this group worked on, we focused heavily on implementing the machine learning model. In turn, this did not give us much time to work on building an interesting interface or way to display the results. Comparatively, we began working on the interface for this project shortly

after establishing a project design. This allowed us time to experiment with different features, explore what types of visuals would be interesting, and get a good understanding of the capabilities of the Dash within Python for building a web interface. This approach helped reduce stress and give a more balanced and professional looking project. This is a strategy that will certainly be helpful in future projects in a real-world setting.

We discovered how difficult data collection can be while working on data science projects. Often in classroom settings, we utilize Kaggle or similar sites to retrieve datasets that can be used to practice our machine learning skills. However, in the real world, the process of finding useful datasets becomes much more complex. For certain data types, such as traffic data, it was ultimately much more of a process for us to find datasets that we could work with properly. Highly-sought out datasets like this often end up setting low limits in terms of how much of the data we can actually use. Additionally, sometimes they are only available to a certain point, which can be difficult when we need to use historical data to find patterns or averages that can be used within our projects.

Overall, we learned a great deal about project management, the full scope of a data science project, and the importance of proper collaborating techniques. We look forward to taking this knowledge with us in our future endeavors to be well-rounded and impactful data scientists.

Citations

Vision Zero. Boston.gov. (2018, March 1). <https://www.boston.gov/departments/transportation/vision-zero>

Moosavi, S. (2023, May 28). US accidents (2016 - 2023). Kaggle. <https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents>

MassDOT: Crash Data Portal. (n.d.). <https://apps.impact.dot.state.ma.us/cdp/home>

Maps and location technology. TomTom. (n.d.). <https://www.tomtom.com/>